My Security: An interactive search engine for cybersecurity

Nan Sun, Shang Gao, Leo Yu Zhang
Deakin University
mzz, shang.gao, leo.zhang@deakin.edu.au

Seyit Camtepe
Data61, CSIRO
seyit.camtepe@data61.csiro.au

Jun Zhang, Yang Xiang
Swinburne University of Technology
junzhang, yxiang@swin.edu.au

Abstract

Large volumes of Cybersecurity-related data is generated every day from various sources at high speed to adapt to the fast-evolving landscape of cybersecurity. It drives the emergence of challenges such as the efficient gathering of in-demand information from unstructured and heterogeneous data sources. After collecting sufficient data, it is hard for users to understand the message hidden behind without adequate security domain knowledge. To help address this problem, in this paper, we present My Security, an innovative search engine for gathering, managing, and understanding cybersecurity-related data. My Security is based on a novel indexing approach that stores both the information of data sources (e.g., publication date, authorship) and the pragmatics messages, including security category (e.g., ransomware, data breach) and corresponding security components (e.g., time of the event, impacted systems). With the established index, users can retrieve cybersecurity information through comprehensive approaches. Fetched results are provided with interpretations leveraged from pragmatics indexing. Additional data mining and visualization techniques enhance the interactivity of My Security by presenting the retrieved results in a clear and comprehensible manner with cybersecurity expertise. It is demonstrated that My Security is efficient at satisfying users' requirements for searching security data and helping people gain better insights into cybersecurity.

1. Introduction

In recent years, potential threats associated with cybersecurity have increased at an alarming rate. It impacts not only the privacy of sensitive personal information but also the security of governments, industries and enterprises in the Internet-connected world. Research communities and cybersecurity sectors are continuously developing and deploying solutions to improve cyber resilience. The knowledge on cyber threat intelligence, such as the mechanisms, indicators and actionable strategies, are effectively shared and exchanged. Such information is valuable to gain insight into the rapidly evolving cybersecurity landscape, not only for individual users but also for organizations. Besides, the indicators hidden in cybersecurity data are extracted and utilized to detect, predict and monitor cybersecurity incidents and attacks [1].

Challenges: With large volumes of cybersecurity data being produced at high speed from different sources, challenges are raised, e.g., how to retrieve cybersecurity-related data that meets specific requirements efficiently. Employing search engines is one of the solutions. As one of the most popular tools for information retrieval, it retrieves results from enormous number of web pages based on users' query. There have been significant efforts in designing and developing advanced search engines. However, little work has been dedicated to the field of cybersecurity [2, 3]. Shodan is the world's first security-related search engine for Internet-connected devices [4]. It scans the Internet to find open ports on a given IP address. Although Shodan provides comprehensive information about the scanned devices, it is desirable to receive information more understandably for people without enough cybersecurity background. A customized Google search engine [3] can index cybersecurity-related websites and query them. Compared with the general-purpose Google search engine, the custom search engine only returns links of websites containing query keywords rather than fine-tuned responses via multiple interactions with users.

My Security: In this paper, we propose and implement a cybersecurity search engine to bridge the gap identified above. Intuitively, the information hidden behind data received by users is decided not by the number of results returned, but by the perceived level of their understanding. Our search engine, named as My Security, aims to provide customized results to users in an understandable and professional manner based on users' query. At the back end, heterogeneous data

URI: https://hdl.handle.net/10125/71369
978-0-9981331-4-0

that includes web pages and also other cyber threat intelligence from multiple sources is processed and indexed. The texts containing context, description and other textual contents are tagged with security entities that interpret the security events by word embedding neural network models. At the front end, the retrieved results are presented to users based on the pragmatics index. In virtue of data analytics and visualization approaches, our search engine enables the output of search results to be customized case-by-case according to users' demand.

Contribution: My Security is an interactive search engine for retrieving information based on users' query in the cybersecurity domain. It addresses the emerging challenge on how to obtain and analyze data effectively from massive open-source cybersecurity data. A novel indexing approach is proposed and implemented that retains the desirable characteristics of conventional indexes and combines with the pragmatics instilled from textual information. Based on the index, results are presented to users with specialized, scalable and understandable features. To the best of our knowledge, no existing work has reported the same functionality. The search engine is evaluated based on modules implemented. The result shows My Security can effectively present retrieved results to users and provide better insights into cybersecurity as well as actionable mitigation strategies.

2. My Security: Design and Implementation

Obtaining the required information is still a time- and resource-consuming task. In our system, security information is automatically extracted and indexed by utilizing conventional technological terms before information retrieval, which makes data more manageable. Based on the index, information filtering and retrieval can be conducted efficiently, making it possible to deliver accurate, rapid and straightforward results to users with data mining and visualization techniques applied. In this section, we present the high-level design of My Security system and describe how its flexible architecture is implemented.

2.1. Data tagging and indexing

Cybersecurity incident detection and prediction are both driven by data and relied on data [1]. The interpretation and exhibition of large volumes of time-oriented data are essential to address security challenges (see Section 4). Figure 1 depicts the indexing process of My Security, including text preprocessor and metadata preprocessor. Heterogeneous data from

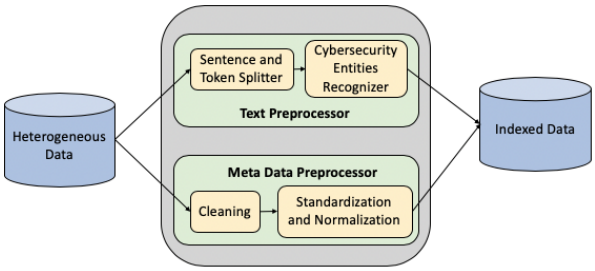


Figure 1. Indexing process of our search engine.

various data sources is filtered and categorized into two main categories: text data that includes sentences and paragraphs of the heterogeneous data; metadata that describes the information of the text data.

Within the text preprocessor module, the cybersecurity entity recognizer is a neural network model [5], working offline to capture entities that are critical to a cybersecurity event. Under the assumption that news and articles relevant to cybersecurity are adequately written, the model is trained on 1000 cybersecurity articles that discuss five main categories of security events, including data breach, phishing, ransomware, vulnerability discovery, and vulnerability patching. Besides extracting the words or phrases that indicate a specific security event category in a sentence, the cybersecurity entity recognizer also recognizes words or phrases that describe the critical elements of a security event, such as time of the event, involved organizations and other specificities of the event. The BIO schema annotates the extracted words and phrases by the entity recognizer, where "O" represents no entity, "B-" for the beginning of entity or "I-" for the continuation of the entity. In total, ten tags are indicating the category of a security event named event nuggets, and 42 tags describing the relevant components of one security event called event arguments. Thereby, the text data is extracted and tagged with corresponding event nuggets and arguments after splitting paragraphs into sentences and sentences into tokens.

Another indexing layer saves metadata that describes the text reserved by the text preprocessor. Examples of metadata involve publication date, information about authorship, source link, frequencies of comments and likes of a data source, etc. After the steps including data cleaning, standardization and normalization, metadata is indexed as structured fields and ready for search.

As shown in Figure 2(a), a Tweet posted on Twitter is used as an example to demonstrate the process of data tagging and indexing. The Tweet flows through the text indexing and metadata indexing modules. The cybersecurity entities mentioned in the Tweet are

(a) A Tweet example from one cybersecurity data source

B-PII	I-PII	B-Databreach	I-Databreach	O	O	B-File	
Email	account	got	compromised	,	please	ignore	strange
I-File	O	B-Person	O	O	O	O	
emails	from	me	with	random	phishing	links	.

(b) Cybersecurity entity extraction and tagging

Figure 2. An example of text data index with annotated security entities.

extracted and respectively tagged as event nuggets and arguments, as shown in Figure 2(b). After that, it is converted to text index with an event ID and tokens with annotations. Information that describes the Tweet including the document ID, the link of the Tweet, author of the post, publication date, the number of comments, the number of retweets and the original text, is indexed as metadata. With all the information captured and indexed, information retrieval can be performed based on this index structure.

2.2. Security information retrieval

For each cybersecurity data record, the entity recognizer identifies each sentence within that record that is likely to include a cybersecurity event and picks up the event nuggets and arguments before text indexing. Together with metadata indexing, we utilize the information retrieval approaches driven by users' demand. Like what a general-purpose search engine provides, a search box is an essential part that provides an interface to query the underlying data. Besides, a selection menu is also offered as a specific function to help users for whom with limited cybersecurity domain knowledge to locate the information they are looking. The options suggested by security experts are provided in the selection menu as an alternative approach to the traditional search box. The directly retrieved results include hyperlinks, supporting users for further navigation. Other information, such as the pragmatics of security events or the publication date stored in metadata index, is retrieved together with the hyperlinks and presented as structured data in a standardized format.

As mentioned, the usefulness of retrievals is determined by how much information users understand instead of results amount. Hence, we further apply data analytics associated with integrated visualization techniques to model, compare and summarize the results. Each result is presented more understandably. In contrast to the batch results, a whole picture of retrieved information is produced for users based on their demands. The detailed information retrieval procedure and the factors associated with data analytics and visualization are elaborated through the system demonstration in Section 3.

2.3. System prototype implementation

My Security prototype is implemented[1] using Shiny [6], which is a robust web framework for building interactive web applications. Three datasets are utilized in our study to establish the prototype: 1000 security news articles [7] that mention five security events and are annotated by experienced security experts; vulnerability archives collected from authoritative vulnerability database [8] and tweets from 2015 to 2020 that mention security keywords and retrieved by a Python package called Twitterscraper [9]. It is worth mentioning that My Security is extendable for both including customized datasets and boosting more functions, such as applying additional data analytics and visualization techniques to satisfy users requirements for specific scenarios.

3. System demonstration and evaluation

In this section, we amplify how My Security works based on the module through use cases. Besides, each module is elaborated and further evaluated with examples of cybersecurity scenarios.

3.1. Security trend

The following use cases demonstrate two typical usage scenarios of the first module of our search engine named "cybersecurity trend". Based on the hypothesis that the discussion frequency of a security issue reflects the impact of that issue, the cybersecurity trend model is designed to observe the complete picture of a set of security data.

Use case 1: *Tom, a security analyst, was provided with a mass of internal cybersecurity data collected at different time ranges and about various security attacks on his first day of work. He planned to get an understanding of these data and the general security situation on the company as quickly as possible. However, he found it was hard to narrow down the search area and then locate suspicious security activities. It would be helpful for Tom to get an impressive idea about the security situation in his new shift.*

[1] https://nansun77.shinyapps.io/cyberSecuritySearchEngine/

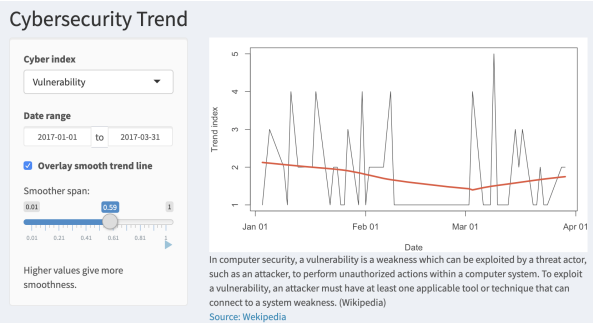


Figure 3. An example of search result by security trend module - the overview of vulnerability trend in the first quarter of 2017.

Use case 2: *Tom has gained experience and capability gradually after working for a few years. He was expected to generate an annual cybersecurity report for his company. Although Tom got into the details through daily observation, it would be beneficial to show Tom an overall trend of cybersecurity at the diverse security index level (e.g., attack categories).*

Demonstration: Figure 3 shows an output example of cybersecurity trend, which pictures the overview of data relevant to security issue "vulnerability" during the first quarter of 2017. To better facilitate users to explore the potential trend hidden behind data, a smoother is added to polish the tendency. Users can modify the smoother span to control the percentage of points in the plot. The higher value on the smoother span, the smoother curve is plotted. The smoother is developed based on the robust locally-weighted polynomial regression curve proposed by Cleveland [10]. Usually, radical changes like sharp rise can give warnings to users and reflect the severity of a specific security issue. Given the above use cases, this module provides an approach to narrow down search area, motivating users to investigate more in-depth into the irregular period and then locating suspicious activities. Also, users can limn the data in a selected period based on a chosen cybersecurity index.

Evaluation: To evaluate the effectiveness and efficiency of the module, we use the trend of security vulnerabilities as an example. Given the selected index being vulnerability, we collect the Twitter data ranging from 2014 to 2018 and containing the Common Vulnerabilities and Exposures (CVE) ID in a format of "CVE prefix-year-arbitrary digits" (e.g., CVE-2018-0001). By mapping each vulnerability to the corresponding weakness with Common Weakness Enumeration (CWE) ID, we rank the number of the most frequently discussed weakness category and visualize the top five most-discussed weaknesses on the heat map.

As shown in Figure 4, CWE−119 (Buffer Errors)

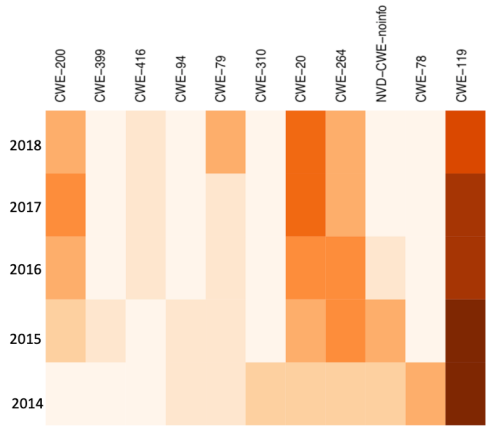
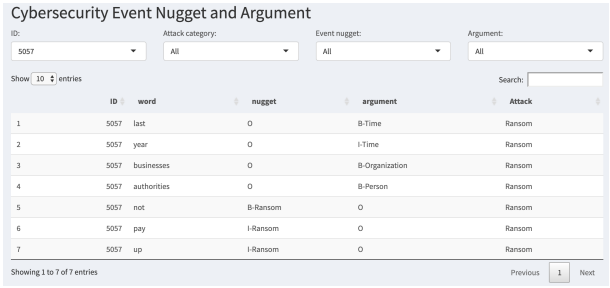


Figure 4. Trend of discussed weaknesses across five years on heat map (darker color represents more relevant data found).

is always the most often noticed weakness in the last five years. This kind of weakness usually happens under the condition that software can read or write a memory location outside the boundary of a buffer when performing a particular operation on memory. It is worth noting that CWE−310 (Cryptographic Issues) reached a peak in 2014 and then eased in the next four years, due to many vulnerabilities related to SSL server certificates were discovered in Android applications in 2014. Another feature is the CWE−200 (Information Leak/Disclosure), which should be highly prioritized due to the nearly doubled number since 2015. According to the trend analysis of Annual Cybersecurity Report [11] from NTT Secure Platform Laboratories, in 2015, the top five critical risk weaknesses were CWE−119 (Buffer Error), CWE−79 (Cross-site Scripting), CWE−200 (Information Leak/ Disclosure), CWE−264 (Permissions, Privileges and Access Control) and CWE−20 (Input Validation), which precisely align with the trend plotted in Figure 4. This result demonstrates that the frequency of discussion on a specific security issue can indeed reflect the trends. Also, observing cybersecurity trends facilitates the research communities and industries to review and reflect on the historical security threats.

3.2. Event pragmatics

During the process of data indexing, the semantic components that depict the critical information about a cybersecurity event are automatically extracted. To aid in users understanding of the cybersecurity data, the second module named "event pragmatics" provides a way to grasp contextual information of an event, such as the impacted organization, event location, date and other messages.



(a) Event nuggets and arguments

(b) Event description

Figure 5. An example of event pragmatics with annotated security entities

Use case 3: *Tom observed a suspicious security activity at work. However, due to his lack of security domain knowledge, he could not interpret it. In this scenario, support should be given to help Tom understand the crucial entities (e.g., attack type, impacted system, etc.) existing in the suspicious activity.*

Use case 4: *Tom was provided with a bunch of reports that record the historical cybersecurity events. Tom didn't have enough time and energy to go through all of them and learn from each event. In this case, if phrases or words that describe the critical entities of each security event are automatically extracted and annotated, Tom can effectively learn from these reports as well as the explainable annotation brought by the pragmatics.*

Demonstration: We leverage a customized neural network model that incorporates linguistic features and word embedding proposed in our previous work [5] to conduct the cybersecurity entity recognition. For each data record that includes text, our entity recognizer automatically labels the words and phrases that represent the event category as event nuggets and annotates the corresponding words and phrases that describe the critical details of the security event as event arguments. Figure 5 demonstrates an example of nuggets and arguments recognized for the event with ID 5057. Given the plain description of this event shown in Figure 5(b), users may not interpret the security issue hidden behind the text. The event pragmatics module directly picks up the entities relevant to security and presents the simplified and understandable results to users. Users can easily understand the specific event category that the data belongs to and the related details that are strongly associated with this security event. As one of the examples of event pragmatics search details presented in Figure 5(a), the event with ID 5057 records a ransomware attack, together with the event occurrence time, the organization and people involved.

Evaluation: The event pragmatics module automatically extracts the entities that are indicators of the event category and other details describing the event to output understandable search results. There has been some research work focusing on collecting information on one particular type of cyber attack. For instance, Work [12] specialized in picking up the four-dimensional key security factors from the text when it comes to ransom. Besides, work [13] concentrated on extracting security components related to malware. Moreover, some existing work [14, 15] conducted a proof-of-concept to demonstrate that their approaches could cover more than one security event type. However, the security details hidden in the text were not specified, making it hard to facilitate users' further understanding of the security components. From the quantitative measurement shown in Table 1, it can be seen that our approach covers the most extensive scope and supports the complexity of cybersecurity event types.

3.3. Event visualization

The event pragmatics module gives hints on how to interpret data on the basis of one single event. In the process of the daily management of cybersecurity activities, people face tremendous amounts of data. The "event visualization" module provides an interface that shows the relationship of events by picturing an overall view of them.

Use case 5: *After performing event pragmatics operations, Tom can easily unscramble cybersecurity events recorded in the reports. Tom was intended to learn from historical processes (e.g., event impact or mitigation strategies) when he observed a new activity that appeared to be similar to an existing event. It is desirable if there is a visualization tool that graphically presents the relationships among events rather than manually search.*

Table 1. Comparison of security entity recognition approaches.

	Data Breach	Phishing	Ransom	Discover Vulnerability	Patch Vulnerability	DDoS	Account Hijacking	Malware
Khanhdur et al.	0	No	No	No	No	0	0	No
Lim et al.	No	No	No	No	No	No	No	12
Ariffini et al.	No	No	4	No	No	No	No	No
Qiu et al.	The event is annotated to "cyber attack" or "other". Event arguments are not specified.							
Ours	15	14	13	10	11	No	No	No

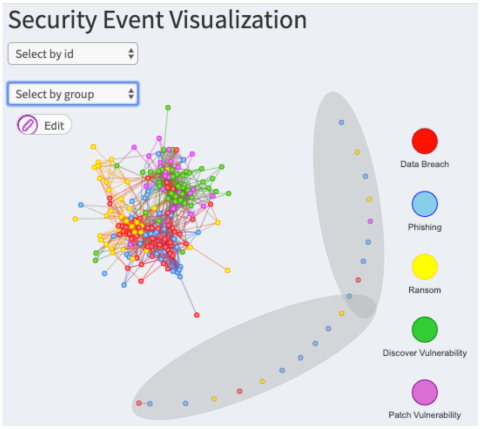


Figure 6. Security event visualization interface with sample events from security categories of data breach, phishing, ransomware and vulnerability.

Demonstration: As demonstrated in Figure 6, each recognized security event is denoted by a node, and the edge between two nodes represents the relationship between the two events. The weight of an edge between two nodes shows the relevance degree between two events. If there are no common properties between two nodes, the weight between the two is 0. If two nodes have one common property, the weight between them will increase by one. The properties come from the argument annotations that are recognized by the event pragmatics model. For example, if two events happen on the same day, but the other features, such as the event location, impacted systems or involved people, are all different, the weight between these two events is 1.

Evaluation: Given a number of security events, the event visualization module offers an effective way to deliver comprehensive information about the possibilities to users. We evaluate the event visualization module by elaborating how it impacts cybersecurity management. On one hand, if the newly observed event is closely connected to other historical security events, users can refer to experience of handling the historical events, such as the mitigation strategies, the impact and indicators of threats. On the other hand, if the observed event has no connection to any existing events, it can be deemed as an outlier (shown in the shadow area in Figure 6). The outlier points that are visualized outside the usual range represent the events that have a lower similarity of characteristics to most of other observed events. In practice, the anomalous events highlighted as outliers need more special attention in case of potential unknown threats.

3.4. Security information search and visualization

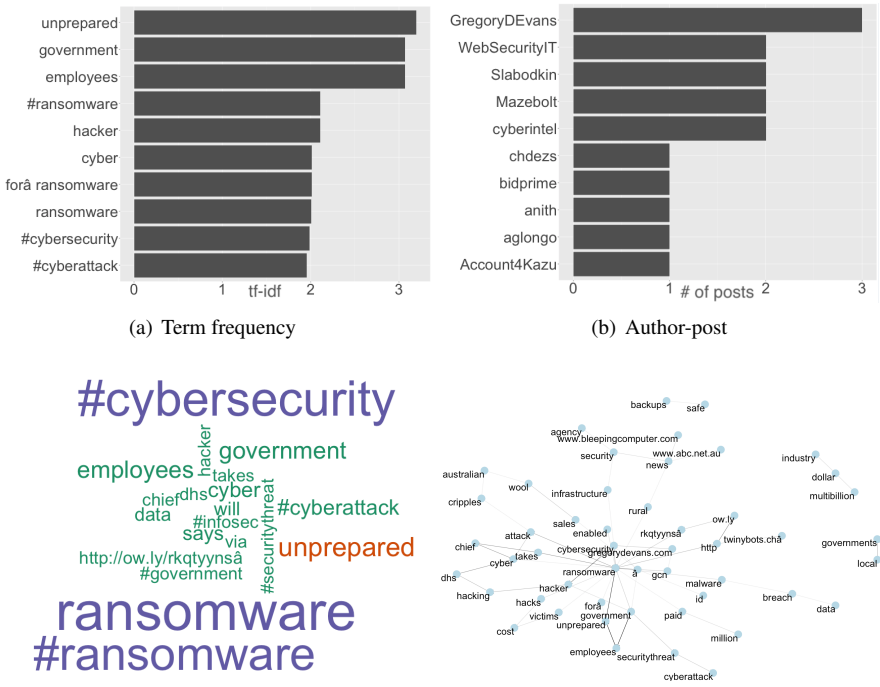
The "security information search and visualization" module behaves like a regular search engine, which implements the basic search functionality. Besides of the general search function, My Security also visualizes the results in various ways facilitated by data analytics techniques.

Use case 6: *Tom generated a hypothesis about a potential threat. He planned to make further investigation on this threat to support or reject the hypothesis by finding more evidence. It would be helpful to trace back to the source data.*

Use case 7: *Tom collected information using keywords in daily work. Heterogeneous data (e.g., papers, reports, etc.) were retrieved containing the keywords. It was expected that the concerned data is obtained in an understandable and organised way rather than title-link result enumeration.*

Demonstration: As the "Basic Search" module in the implemented prototype system, it helps users gather information based on their queries, also delivers the straightforward visualization of data's inherent characteristics and comprehensive summary. We demonstrate the basic search module from the perspective of search options and data mining-based visualization, respectively.

Users can search for specific cybersecurity topics using keywords. The content-based search field supports search by security category, author, union, intersecting and excluding keywords. Besides, the numerical search fields support date range, number of likes and number of comments on data source. The advanced search function is employed to conduct



(a) Term frequency (b) Author-post

(c) Word cloud (d) Bigrams cloud

Figure 7. Examples of visualized results when searching keyword "phishing" with time range from 01/01/2020 to 27/02/2020.

sophisticated queries by combining the text search fields and numerical data fields.

Results are retrieved based on search input. Like results returned by a traditional search engine, users can locate the information source by clicking a website link of a retrieved result (if it exists) to explore further. However, numbers don't convey the information they need to see very efficiently for many people. Visualizing data is often a more effective way to discover patterns in it. My Security integrates the visualization tools that facilitate the understanding and summarization of large volumes of data. The term frequency chart groups terms by the number of mentioned times within the search range and shows their usage frequency on cybersecurity topics. As shown in Figure 7(a), due to the trade-off between effectiveness and simplicity in describing the relevance of the term to a document, TF (Term Frequency) - IDF (Inverse Document Frequency) is adopted to measure term frequency. TF-IDF value increases with the number of term occurrences in the document and decreases with the number of documents where the term appears. Suppose there is a collection of N documents, the function df (document frequency) is defined by the terms in the dictionary and returns the number of documents where the term occurs. For a term-document pair (t, d), the term frequency tf (t, d) is defined as the number of occurrences of the term in the document. Hence, the tf-idf is defined by the pair (t,d) as:

$$tf - idf(t,d) = tf(t,d) \cdot \log \frac{N}{df(t)}.$$

Besides, the number of posts published by a particular source enables the investigation on the quality of data sources [16]. Authors' information and their post number ranking are visualized as shown in Figure 7, which inspires users to evaluate the quality of data and balance the completeness of data sources.

Table 2. Observed frequencies of word X and Y in selected data.

	Has word X	No Word X	Total
Has Word Y	n_{11}	n_{01}	$n_{1.}$
No Word Y	n_{10}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	n

Word clouds (Figure 7(c)) can be generated based on both term frequency and TF-IDF in My Security system. Compared with term frequency bar chart, the word cloud is a novel visual representation of text data. It is used to depict keywords and represent the importance of each keyword with font and color gradient intuitively. Bigrams cloud (Figure 7(d)) visualizes the word pairs that are highly related to each other. It helps users gain an insightful understanding of data. The pairs

of two words are examined on how often they appear together relative to how often they appear separately. Given two words X and Y, coefficient ϕ [17], a common measure for binary correlation, is utilized to calculate the correlation between X and Y. As shown in Table 2, the number of data records where both Word X and Y appear, the number of data records where neither appears is respectively represented as n_{11} and n_{00}. The number of cases where one word appears without the other is denoted as n_{01} or n_{10}. Then the coefficient ϕ is calculated by

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}}.$$

By visualizing the correlations between word pairs, bigrams cloud helps find the other terms that are most or potentially relevant.

Evaluation: We summarize and compare the existing cybersecurity domain search engines with our basic search module from the perspective of functionality, results' presentation and elasticity, as shown in Table 3. Following the instructions of Google Custom Search [3], we build a customized cybersecurity search engine[2]. It covers the top 10 cybersecurity intelligence sources investigated and measured by [16]. Besides, several public cybersecurity websites (such as CVE [8]) are included in the customized cybersecurity search engine.

From the functionality perspective, all the existing cybersecurity domain search engines support cybersecurity information search through the search box. Besides of the essential search function, Shodan and My Security both provide an interface to explore information within featured security categories. Also, the retrieved results in Shodan and My Security are summarized and can be delivered as customized reports. For users with limited security domain knowledge, My Security automatically extracts the pragmatic information, delivering more understandable and explainable results.

When looking into the representation of results, Google customized search engine delivers results in a format of title and link combination. Shodan, as the first search engine for connected devices, manifests search results by taking advantage of structured information on connected devices and several straightforward visualization techniques, such as bar charts and device location representation through a world map. Our search engine not only provides title-link results but also makes use of multiple statistics- and analytics-based visualization tools to present results.

[2] https://cse.google.com/cse?cx=012104931910478407356:vdp4jcckw2x

Table 3. Existing cybersecurity search engines comparison.

		Google Customized Search	Shodan	My Security
Functionality	Search	✓	✓	✓
	Navigation		✓	✓
	Explanation			✓
	Summary		✓	✓
Presentation	Brief & link	✓		✓
	Structured data		✓	✓
	Frequency bar chart		✓	✓
	Analytics based visualization		✓	✓
Elasticity	Open source			✓
	Extendable			✓

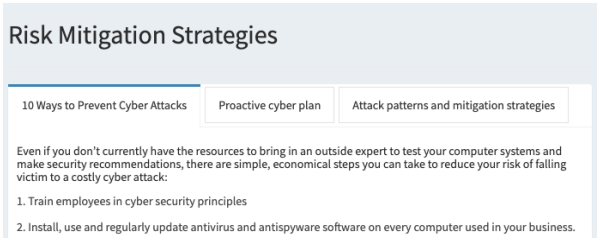


Figure 8. Mitigation strategy interface with three main schemes of security management.

My Security is open source and written in R programming language. It provides a higher degree of elasticity for customized features. The system is extendable to incorporate other public or private datasets, and can easily integrate new data analytics and visualization techniques. In the long term, especially for organizations and enterprises, users can have their own security better enhanced utilizing human argumentation, such as continually motivating employees to empower and update with new security knowledge [18].

3.5. Mitigation strategies

Users, especially these who are on duty of security and risk management, are increasingly seeking security and risk management solutions with capabilities supporting proactive protection, prevention and response. The "Mitigation strategies" module offers three primary schemes of security management, which can be selected, run and tuned by users based on their resources and capabilities.

Use case 8: *Tom confirmed a hypothesis about a potential threat based on the collected evidence. Due to lack of experience with cybersecurity management, Tom felt clueless about the impact of the threat and how to mitigate it. In this case, Tom was expecting to seek out experts' advice and get professionals' help.*

Demonstration: Awareness and education on

cyber threats play an important role in enhancing security for organizations and enterprises, especially for small businesses [19]. According to statistics, 60% of small businesses that were compromised by cyber attacks permanently closed within six months of attacks. To help users reduce the risk of being victimized, our mitigation strategies module lists tips on general practices as well as the ten ways to prevent cyber attacks, aiming at those who are still under the weak sense of cybersecurity.

In addition, we utilize the Cyberplanner [20] developed by Federal Communications Commission (FCC) to create customized cybersecurity plans for individual companies, addressing their specific concerns and needs. Topics in the plan are proposed by cybersecurity experts and can be selected by users based on their demand. It provides customized professional suggestions on mitigation of security attacks. In the meanwhile, the mitigation strategies remind users to prepare in advance before damage was made, promoting proactive policy rather than reactive.

Users can search for risk mitigation strategies through My Security by keywords. For each specific attack pattern, relevant information including the impact of attacks, mitigation strategies and related weaknesses are all enumerated according to community resources Common Attack Pattern Enumeration and Classification [21]. As shown in Figure 8, the mitigation strategies model offers an hierarchical suggestion schema for users in different contexts to apply strategies in practice.

4. Related work

To help security communities and organizations defend against the fast-evolving cybersecurity attacks, many efforts have been made on sharing information on security and threats, vulnerability and incidents. These cybersecurity data sources include evidence-based knowledge, involving indicators, implications and actionable advice. They can be leveraged to contribute to decision making and response to emerging and existing hazard [22]. Large volumes of cyber threat intelligence data have been generated publicly and privately, no matter by number, range or scale. Driven by the increment, research communities and industries have been utilizing different kinds of data sources to improve cyber resilience [1, 23].

There are various types of cybersecurity data sources used in previous studies. Security data directly crawled from web pages provides context as well as meta-information such as authorship, publication date in HTML format, which was broadly used in previous work to discover a specific kind of threat [24][25]. Besides, social media platforms that generate a steady flow of information ensure the innovative strength to contribute to cyber threat intelligence. Previous work made full advantage of social media data to discover indicators of compromises (IOCs) [16], detect malicious mobile applications [26] and find cyber attacks [14][23]. Some authoritative datasets published by government and security sectors make data more reliable. For example, Common Vulnerability and Exposures (CVE) is the world's leading organization to provide vulnerability information to predict real-world vulnerability exploits. Sabotke et al. [27] combined Tweets and CVE information in their datasets to predict whether a vulnerability would be exploited or not.

The investigation on previous work suggests that cybersecurity information can be collected from various data sources. Combining data from multiple sources makes the data more comprehensive, reliable and innovative for cybersecurity operations and risk management activities. Information retrieval is a science of searching for information from text, images or sounds, and also searching for information from metadata that describes data [28]. By means of information retrieval, information can be effectively filtered and selected according to requirements.

A search engine that implements information retrieval in practice is one of the most widely used methods for navigating cyberspace. Usually, a search engine keeps a copy of extensive collections of web pages and related information using URLs as row keys and various aspects of web pages as columns. Search algorithms sort through hundreds of billions of web pages to locate the most relevant results and present them in multiple formats. However, there is little work in the field of specific domain [3][4], primarily cybersecurity domain. From the viewpoint of users, on the one hand, users expect to retrieve precise information quickly and simply. On the other hand, it is desirable to get interpretation support to aid users to understand the information better. We thereby make the ultimate goal of designing a cybersecurity domain search engine. It learns from the existing search algorithms and the interactive features boosted by these search engines. Meanwhile, there is a tradeoff between insufficient cybersecurity domain knowledge and overloaded awaiting ingestion information. By applying abundant data analytics and visualization techniques associated with pragmatics, our search engine is designed to provide direct and efficient results in an easy and understandable fashion.

5. Conclusion

In this paper, we present My Security, a novel search engine designed to specialize in cybersecurity

information retrieval. It retrieves information based on users' queries from heterogeneous data sources. Based on the proposed indexing approach, the pragmatics and meta-information hidden behind data are efficiently recognized and indexed. By extracting the security entities from sentences, search results are delivered with comprehensive interpretations of the security events, including the security category and detailed components of each event. Moreover, My Security leverages assorted data mining and visualization techniques to support users to understand fetched search results interactively. My Security is found to be highly effective and outperforms the existing cybersecurity domain search engines in terms of functionality and elasticity. The prototype system is established based on over 1,000 security news, five years of security Tweets and public cybersecurity databases. Its effectiveness is demonstrated, highlighting the significance of the security domain search engine and advancing users insights in cybersecurity.

References

[1] N. Sun, J. Zhang, P. Rimba, S. Gao, L. Y. Zhang, and Y. Xiang, "Data-driven cybersecurity incident prediction: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1744–1772, 2019.

[2] J. Matherly, "Complete guide to shodan," *Shodan, LLC (2016-02-25)*, vol. 1, 2015.

[3] "Google custom search." `https://cse.google.com/cse/all`, 2019.

[4] "Shodan." `https://www.shodan.io/`, 2019.

[5] N. Sun, J. Zhang, S. Gao, L. Y. Zhang, S. Camtepe, and Y. Xiang, "Cyber information retrieval through pragmatics understanding and visualization," 2020.

[6] W. Chang, J. Cheng, J. Allaire, Y. Xie, J. McPherson, *et al.*, "Shiny: web application framework for R," *R package version*, vol. 1, no. 5, 2017.

[7] T. Satyapanich, T. Finin, and F. Ferraro, "Extracting rich semantic information about cybersecurity events," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 5034–5042, IEEE, 2019.

[8] "Common vulnerabilities and exposures." `http://cve.mitre.org/`, 2018. Accessed on 11/09/2019.

[9] A. Taspinar and L. Schuirmann, "Twitterscraper 0.2. 7: Python package index," 2017.

[10] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

[11] N. S. P. Laboratories, "Annual cybersecurity report." `http://www.ntt.co.jp/sc/media/NTTannual2016_e_web_lock.pdf`, 2016. Accessed on 11/11/2019.

[12] N. Ariffini, A. Zainal, M. A. Maarof, and M. N. Kassim, "Ransomware entities classification with supervised learning for informal text," in *2019 International Conference on Cybersecurity (ICoCSec)*, pp. 86–90, IEEE, 2019.

[13] S. K. Lim, A. O. Muis, W. Lu, and C. H. Ong, "Malwaretextdb: A database for annotated malware articles," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1557–1567, 2017.

[14] R. P. Khandpur, T. Ji, S. Jan, G. Wang, C.-T. Lu, and N. Ramakrishnan, "Crowdsourcing cybersecurity: Cyber attack detection using social media," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1049–1057, 2017.

[15] X. Qiu, X. Lin, and L. Qiu, "Feature representation models for cyber attack event extraction," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW)*, pp. 29–32, IEEE, 2016.

[16] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, "Acing the IOC game: Toward automatic discovery and analysis of open-source cyber threat intelligence," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 755–766, 2016.

[17] H. Cramir, *Mathematical Methods of Statistics*. Princeton U. Press, Princeton, 1946.

[18] J. Groopman, "AI-driven Cybersecurity Teams Are All About Human Augmentation," 2020.

[19] K. F. McCrohan, K. Engel, and J. W. Harvey, "Influence of awareness and training on cyber security," *Journal of Internet Commerce*, vol. 9, no. 1, pp. 23–41, 2010.

[20] F. C. Commission, "Cyberplanner." `https://www.fcc.gov/cyberplanner`, 2020. Accessed on 21/9/2020.

[21] "Common attack pattern enumeration and classification." `https://capec.mitre.org/`, 2018. Accessed on 20/12/2019.

[22] R. McMillan, "Open threat intelligence." `https://www.gartner.com/en/documents/2487216`, 2013.

[23] N. Sun, G. Lin, J. Qiu, and P. Rimba, "Near real-time twitter spam detection with machine learning techniques," *International Journal of Computers and Applications*, pp. 1–11, 2020.

[24] K. Soska and N. Christin, "Automatically detecting vulnerable websites before they turn malicious," in *23rd USENIX Security Symposium*, pp. 625–640, 2014.

[25] K. Borgolte, C. Kruegel, and G. Vigna, "Delta: Automatic identification of unknown web-based infection campaigns," in *Proceedings of the 2013 ACM SIGSAC conference on Computer Communications Security*, pp. 109–120, 2013.

[26] D. Kong, L. Cen, and H. Jin, "Autoreb: Automatically understanding the review-to-behavior fidelity in android applications," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 530–541, 2015.

[27] C. Sabottke, O. Suciu, and T. Dumitraș, "Vulnerability disclosure in the age of social media: Exploiting twitter for predicting real-world exploits," in *24th USENIX Security Symposium*, pp. 1041–1056, 2015.

[28] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.